ABSTRACT OF THE DISCLOSURE

A media capture device has an audio input receptive of user speech relating to a media capture activity in close temporal relation to the media capture activity. A plurality of focused speech recognition lexica respectively relating to media capture activities are stored on the device, and a speech recognizer recognizes the user speech based on a selected one of the focused speech recognition lexica. A media tagger tags captured media with generated speech recognition text, and a media annotator annotates the captured media with a sample of the user speech that is suitable for input to a speech recognizer. Tagging and annotating are based on close temporal relation between receipt of the user speech and capture of the captured media. Annotations may be converted to tags during post processing, employed to edit a lexicon using letter-to-sound rules and spelled word input, or matched directly to speech to retrieve captured media.